DOCUMENT RESUME

ED 292 887 TM 011 350

AUTHOR Mislevy, Robert J.
TITLE Item Response Theory in Educational Assessment.
PUB DATE Apr 88
NOTE 13p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 5-9, 1988).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Educational Assessment; Evaluation Methods; *Item Analysis; *Latent Trait Theory; National Surveys; Scoring; Statistical Analysis; Test Theory
IDENTIFIERS *National Assessment of Educational Progress

ABSTRACT

Large-scale educational assessments differ from familiar educational measurements by attempting to provide information about the levels and natures of skills in populations rather than in individuals. That the distinct purposes of assessment require different methodologies than individual measurement was recognized by the development of multiple-matrix sampling designs. Similar changes in thinking are necessary when it comes to applying item response theory (IRT) measurement models. This paper outlines three challenges to IRT in the assessment setting, and discusses the responses to them taken in the recent surveys of reading, mathematics, and science undertaken by the National Assessment of Educational Progress. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

ROBERT J. MISLEVY

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ITEM RESPONSE THEORY IN EDUCATIONAL ASSESSMENT

Robert J. Mislevy
Educational Testing Service

Abstract

Large-scale educational assessments differ from familiar educational measurements by attempting to provide information about the levels and natures of skills in populations rather than in individuals. That the distinct purposes of assessment require different methodologies than individual measurement was recognized by the development of multiple-matrix sampling designs. Similar changes in thinking are necessary when it comes to applying item response theory (IRT) measurement models. This paper outlines three challenges to IRT in the assessment setting, and discusses the responses to them taken in the recent surveys of reading, mathematics, and science undertaken by the National Assessment of Educational Progress.

Presented at the Annual Meeting of the American Educational Research Association in New Orleans, April 5-9, 1988.

INTRODUCTION

Large-scale educational assessment in the United States, as exemplified by the National Assessment of Educational Progress (NAEP), is intended to provide information about what young people know and can do--in the aggregate. It is not intended to provide precise statements about the capabilities of individuals, as are required for the purposes of diagnosis, selection, or placement. The latter purposes are served by the more familiar tests and methods of educational measurement of individuals, or educational measurement, for short.

It is not surprising that the standard statistical tools, models, and sampling designs that evolved as optimal for educational measurement are not optimal for the purposes of assessment. Multiple-matrix item sampling designs (Lord, 1962; Sirotnik and Wellington, 1977) demonstrate the improvements to efficiency attainable by using sampling plans keyed to measuring groups rather than individuals; in effecient assessment designs, individual persons are asministered very few items in any reporting area. In the same way, the techniques and the models of psychometric measurement theory must be rethought if they are to be used profitably in assessment.

After a brief review of the potential advantages of using item response theory (IRT) measurement models in assessment, this paper outlines three challenges this new setting presents. The responses to these challenges that were developed for the analysis of the most recent NAEP reading, mathematics, and science surveys are then discussed.

EDUCATIONAL ASSESSMENT

As stated above, an educational assessment like NAEP is intended to provide information about the nature and the levels of skills in populations rather than of individuals. Since the skills of interest are the same as those of interest in educational measurement, however, both endeavors proceed by asking students to solve educationally relevant problems. In isolation,

there is no way to tell whether an exercise came from an assessment or a test for individuals.

An assessment is not a test in the usual sense of the term, though. The distinction lies not so much in the nature of the items, but in the way responses to them are summarized and analyzed. Now any method of summarizing data, whether it is purely empirical like an average or model-based like an IRT ability estimate, collapses information over some patterns in order to highlight others--presumably the patterns of interest. In educational measurement, the summarization is carried out over items within individuals, in order to provide for accurate inferences about specific persons. In assessment, the most important summarization is over individuals, to provide for accurate inferences about groups. In fact, the original plan for reporting NAEP results was to give only group-level estimates of percent-correct for items ("item-p's"); no summarization whatsoever over items was envisaged.

While this level of detail is useful to curriculum specialists, it soon become apparent that single-item reporting was too cumbersome for communicating NAEP results to policy-makers and to the public at large. NAEP, at that time administered by the Education Commision of the States (ECS), quickly moved to reporting average percents-correct over sets of related items.

Average item-p's improved communication to be sure, but a limitation of this approach surfaced over time as NAEP learning areas were assessed the second, then the third time. The problem is that to retain their meaning, average item-p's must be defined on a constant set of items. As old items were released from the assessment and new ones were rotated in, the sets of items upon which NAEP could report trends or compare ages diminished rapidly. When the NAEP contract came up for renewal in 1981, the time was ripe for a new reporting strategy--one that would allow results to be reported on the same scale even though item pools evolved over time. With the exception of ECS, every bidder for the new NAEP

contract proposed the use of item response theory in one form or another.

ITEM RESPONSE THEORY

The attractive feature of IRT for educational assessment was its capability, proven in educational measurement, to provide scores on the same scale from responses to different subsets of test items from the same item pool. IRT was was first introduced for measuring the psychological traits or educational proficiencies of individuals, and available procedures had evolved to suit individual measurement data, but the underlying models seemed equally appropriate for assessment.

The main idea of IRT is that the probability of a correct response to a given item from a given person can be modeled as a function of parameters for that person and parameters for that item. To express an IRT model more formally, let θ denote examine proficiency and let β_j express how the probability of a correct response to Item j depends on θ; then

$$P_j(\theta) = P(x_j=1|\theta,\beta j) = f_j(\theta,\beta_j) \ , \quad (1)$$

where $x_j=1$ denotes a correct response and $x_j=0$, an incorrect one. The Rasch model, for example, posits that β_j is a scalar and that $f_j(\theta,\beta_j) = \exp(\theta-\beta_j)/[1+\exp(\theta-\beta_j)]$; i.e., the log of the odds of a correct response, or the logit, is just the linear function $\theta-\beta_j$.

Under the usual IRT assumption of conditional or local independence, the probability of a response vector $x = (x_1,\ldots,x_n)$ given θ is the product of terms built up from (1):

$$P(x|\theta,\beta_1\ldots\beta_n) = \prod_j P_j(\theta)^{x_j}[1-P_j(\theta)]^{1-x_j} \quad (2)$$

After the observation of a response vector for any subset of items whose parameters are known (in practice, well-estimated rather than known), (2) can be interpreted as a likelihood

function and supports inference about θ. If the model is correct, then, one can make inferences about θ from responses to different items. The validity of such inferences in practical applications depends on the extent to which model assumptions are met. In particular, it is typically assumed that...

- o A single θ variable explains all systematic variation among examinees.
- o Item parameters are invariant over examinee groups.
- o Item parameters are invariant over time.

An IRT model, like any other model, is a simplification of reality. No one expects its assumptions to hold perfectly--nor need they for the model to be useful. In the assessment context, the primary purpose for IRT is to summarize population trends and subpopulation differences in average performance on sets of similar items. But this deceptively simple goal cannot be accomplished by applying familiar rules of thumb and standard analytic procedures designed for IRT in educational measurement (Goldstein, 1979). The following sections discuss in turn three challenges that must be met in applying IRT models to assessment data.

CHALLENGE: Items change at different rates, even different directions, over time.

An IRT model over a set of items places strong restrictions upon the relative changes in performance from one item to the next that can occur for individuals or for groups. It is unrealistic to expect these restrictions to be met perfectly in practice (Goldstein, 1979). Items become easier or harder over time for a number of reasons, including curricular emphasis, secular trends, and security breaches, and these factors will differ in their impact from one item to the next.

RESPONSE #1: Limit the breadth of scales.

Experience in educational measurement shows that IRT models generally fit better when applied to more narrowly defined domains of items. For the problem of characterizing changes over time, we wants scales defined sufficiently narrowly that major changes in curricula and society to have similar effects on all items in a scale (Bock, Mislevy, and Woodson, 1982).

Facing this problem without the benefit of previous experience with IRT in the assessment context, the California Assessment Program (CAP) defined IRT scales very narrowly; e.g., scaling in 25 distinct "indivisible curricular elements" for Grade 3 Language Skills. This number of scales provided schools with detailed feedback on the effects of curricular innovation and change. Subsequent analyses of trend data revealed that (1) item parameter estimates were stable over time, thus confirming the hope of capturing change in scale parameters rather than losing it as model misfit; and (2) clusters of, say, four to eight elements were highly correlated and showed extremely similar patterns of change over time. While the element-level detail remains useful to school personnel, higher-level policy-makers obtain sufficient detail in aggregations over subsets of closely-related elements.

Thus, experience with IRT in educational measurement suggests that one scale is almost certainly not enough to capture patterns of change in a learning area, while CAP's experience with IRT in assessment suggests that twenty-five is more than enough for NAEP's purposes. Consequently, NAEP has employed from four to six subscales in the major learning areas of literacy, mathematics, and science. These areas have been delineated with the assistance of learning area committees, groups comprised of learning-area specialists who can identify major conceptual and curricular fault lines among the skills and the content to be assessed.

RESPONSE #2: Content balancing and "rolling calibration" of items.

The first response described above guarantees that differential patterns of change will be distinguished among key curricular subdivisions within a learning area. Each item has its own story to tell, though, and the remaining unmodelable differential changes among items must still be dealt with. NAEP's second response is based on results about the application of unidimensional IRT models to multidimensional data. Knowing what IRT models pick up when they are incorrectly specified, NAEP designs assessment instruments and analysis plans under which IRT models highlight salient patterns and are resistent against idiosyncratic ones.

The perfect fit of an IRT model would have the following implications in the context of repeated assessments. Item parameter estimates based on data from each time point separately would agree, aside from estimation error, up to a linear transformation. The change in all items could be explained by a change in a single underlying θ distribution. An estimate of the change would have the same expected value no matter which subset of items it were based on, so considerations of item content would not be required.

In practice, of course, the model does not fit perfectly. Simulation studies have shown that when the true model is multidimensional with a single dominant dimension, a unidimensional model picks up the dominant dimension--a weighted average of sorts. Just what that dimension will be depends largely upon the numbers of items tapping the various skills. Once one recognizes that a weighted average will be used, it becomes clear that the weights should be determined rationally rather than by coincidence. In practical terms, item specs count. In NAEP, items are written to detailed specifications provided by learning area committees; assessment items appear in prespecified proportions that reflect the committees' views of their importance. To the extent that the model <u>does</u> fit a scaled area,

this content balancing is irrelevant. To the degree the model does not fit, the committee's importance weightings influence the resulting composite weightings implicit in θ.

When the model does not fit, different estimates of change could result from different subsets of items in a scale. The resulting overall estimate of change is again a weighted average over items that appear at both time points. For the same reasons of representativeness discussed in the preceeding paragraph, this set of "change" items should reflect the composition of the pool. NAEP also uses a graphical technique to ensure that this resulting change estimate is representative: estimates of item parameters from adjacent time points are plotted. If the spread is moderate, the average change over all items included in both assessments is the reported measure of change, and the average change is a good approximation of the change of each specific item. But outlying items are those whose changes depart from typical changes--ones which are suspect of security breaches or idiosyncratic effects. Such items, when they are found, can be treated as if they were separate items in the two time points rather than as the same item.

CHALLENGE: **With many scales and efficient sampling design, it is not possible to get good estimates of θ for each examinee.**

In standard applications of IRT in educational measurement, each examinee is administered enough items to permit a good estimate $\hat{\theta}$ of his θ value. Such estimates are the basis of subsequent placement or selection decisions, and are often used in analyses of distributions of θ or of relationships between θ and collateral variables. Assessment's use of several scales (as advocated above) and its severe time limitations on examinee time preclude the estimation of good point estimates for individuals.

RESPONSE: **Use estimation procedures that allow inferences directly from response patterns to structural parameters.**

The computation of $\hat{\theta}$s as intermediate stage in estimating item parameters and population characteristics is neither necessary in assessment, nor is it desirable. When population characteristics are the object of inference, it is more efficient to employ designs that gather relatively sparse data from individuals (Lord, 1962), and more accurate to use analytic procedures that estimate population parameters directly rather than from intermediate estimates (Mislevy, 1984).

Item parameter estimation from sparse examinee-level data is accomplished by NAEP with Mislevy and Bock's (1983) BILOG computer program. This procedure, based on Bock and Aitkin's (1981) marginal maximum likelihood solution for item parameters, expresses what is known about individual examinees in terms of posterior distributions across the entire θ scale rather than as point estimates. The same ideas can be applied to estimate population characteristics, and a number of examples appear in the recent psychometric literature for doing so--for simple random samples of examinees, however, rather than for the complex stratified and clustered examinee-sampling designs like those of NAEP.

CHALLENGE: **Integrate the IRT model for within-person sampling of item responses with a complex survey-sampling model.**

Suppose that θ values were observed without error for respondents selected in a complex sampling design. We could compute a statistic $t(\underline{\theta},\underline{y})$ to estimate a population quantity $T(\underline{\theta},\underline{Y})$, where y is background data, and a corresponding variance estimate $U(\underline{\theta},\underline{y})$. U expresses uncertainty due to the sampling of subjects. But θ cannot be observed directly even for sampled examinees; only item responses x, which depend stochastically upon θ, can be observed. A way must be found to approximate t from what is observed (item responses x and background variables y) and

to quantify the increase in uncertainty about T due to the latency of θ.

RESPONSE: **Employ Rubin's "multiple imputations" procedures for missing data in surveys.**

NAEP's "plausible values" approach to estimating θ distributions is based on Rubin's (1987) multiple imputations procedures for missing data in sample surveys. The idea is to compute a predictive distribution for each missing value, then perform computations with data sets in which missing values are filled in with random draws from their predictive distributions. The key to its implementation in NAEP is to think of θ as a variable whose value is missing for each respondent. In particular, one can approximate t by its (conditional) expected value:

$$t^* = E[t(\underline{\theta},\underline{y})|\underline{x},\underline{y}]$$

$$= \int t(\underline{\theta},\underline{y})\ p(\underline{\theta}|\underline{x},\underline{y})\ d\underline{\theta}\ ,$$

where

$$p(\underline{\theta}|\underline{x},\underline{y}) \propto \prod_i P(x_i|\theta)\ g(\theta|y_i)\ . \qquad (3)$$

Equation 3 is the product, respondent by respondent, of the conditional distributions of their unobservable θ values, given what was observed from each of them--namely, their item responses x and their background variables y. The first term in the products on the right side of (3), or $P(x_i|\theta)$, is given by the IRT model. In NAEP at present, its parameters are estimated with BILOG, and treated as known in subsequent computations. The second term, $g(\theta|y_i)$, is the conditional distribution of θ given background variables. Its parameters are estimated by the method given in Mislevy (1985), then presently treated as known. (Planned refinements of NAEP IRT procedures call for the

uncertainty associated with item parameters and conditional distribution parameters to be taken into account.)

Closed form solutions of (3) are not generally available. One can approximate t^* by taking a random draw (a "plausible value") θ^*_{mi} from each $p(\theta|x_i,y_i)$--this yields a filled-in dataset--then computing t as if θ^*_ms were θs. Call the result t^*_m. Repeating the entire process M times gives M different estimates of t^*. Their average is final estimate t^* of t.

The variance associated with t^* as an estimate of T is the sum of two components: an estimate U^* of U similarly obtained, plus (M+1)/M times the variance of the t^*_ms. The first component quantifies uncertainty due to the sampling of subjects; the second, uncertainty due to not observing θ values directly.

SUMMARY

- IRT is used in NAEP as a data-reduction tool, not as a psychological model.

- Because the IRT model is not accepted at face value, we must gather and analyze data in a way that minimizes errors due to misspecification.
  - > Scale in areas more narrowly-defined than in individual measurement.
  - > Content balance in item selection.
  - > Recalibrate items periodically if they show idiosyncratic drift.

- Employ estimation procedures unlike those usually employed with IRT in individual measurement.
  - > Bypass estimation for individuals.
  - > Use Rubin's multiple imputations to dovetail the IRT model (for responses within persons) with the examinee-sampling model (for persons from the population).

REFERENCES

Bock, R.D., and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. **Psychometrika**, 46, 443-459.

Bock, R.D., Mislevy, R.J., and Woodson, C.E.M. (1982). The next stage in educational assessment. **Educational Researcher**, 11, 4-11.

Goldstein, H. (1979). Consequences of using the Rasch model for educational assessment. **British Educational Research Journal**, 5, 211-220.

Lord, F.M. (1962). Estimating norms by item sampling. **Educational and Psychological Measurement**, 22, 259-267.

Mislevy, R.J. (1984). Estimating latent distributions. **Psychometrika**, 49, 359-381.

Mislevy, R.J. (1985). Estimation of latent group effects. **Journal of the American Statistical Association**, 80, 993-997.

Mislevy, R.J., and Bock, R.D. (1983). **BILOG: Item Analysis and Test Scoring with Binary logistic models** [computer program]. Mooresville, IN: Scientific Software, Inc.

Rubin, D.B. (1987) **Multiple Imputation for Nonresponse in Surveys**. New York: Wiley.

Sirotnik, K., and Wellington, R. (1977). Incidence sampling: An integrated theory for "matrix sampling." **Journal of Educational Measurement**, 14, 343-399.